METHOD AND SYSTEM OF CORRECTING SPECTRAL DEFORMATIONS IN THE VOICE, INTRODUCED BY A COMMUNICATION NETWORK.

BACKGROUND OF THE INVENTION

Field of the invention

The invention concerns a method for the multireference correction of voice spectral deformations introduced by a communication network. It also concerns a system for implementing the method.

The aim of the present invention is to improve the quality of the speech transmitted over communication networks, by offering means for correcting the spectral deformations of the speech signal, deformations caused by various links in the network transmission chain.

The description which is given of this hereinafter explicitly makes reference to the transmission of speech over "conventional" (that is to say cabled) telephone lines, but also applies to any type of communication network (fixed, mobile or other) introducing spectral deformations into the signal, the parameters taken as a reference for specifying the network having to be modified according to the network.

Description of prior art

The various deformations encountered in the case of the switched telephone network (STN) will be stated below.

1.1. Degradations in the timbre of the voice on

the STN network:

Figure 1 depicts a diagram of an STN connection. The speech emitted by a speaker is transmitted by a sending terminal 10, is transported by the subscriber line 20, undergoes an analogue to digital conversion 30 (law A), transmitted by the digital network 40, undergoes a digital (law A) to analogue conversion 50, is transmitted by the subscriber link 60, and passes through the receiving terminal 70 in order finally to be received by the destination person.

Each speaker is connected by an analogue line (twisted pair) to the closest telephone exchange. This is a base band analogue transmission referenced 1 and 3 in Figure 1. The connection between the exchanges follows an entirely digital network. The spectrum of the voice is affected by two types of distortion during the analogue transmission of the base band signal.

The first type of distortion is the bandwidth filtering of the terminals and the points of access to the digital part of the network. The typical characteristics of this filtering are described by UIT-T under the name "intermediate reference system" (IRS) (UIT-T, Recommendation P.48, 1988). These frequency characteristics, resulting from measurements made during the 1970s, are tending however to become obsolete. This is why the UIT-T has recommended since 1996 using a "modified" IRS (UIT-T, Recommendation P.830, 1996), the nominal characteristic of which is depicted in Figure 2 for the transmission part and in Figure 3 for the receiving part. Between 200 and 3400

Hz, the tolerance is ± 2.5 dB; below 200 Hz, the decrease in the characteristic of the global system must be at least 15 dB per octave. The transmission and reception parts of the IRS are called respectively, according to the UIT-T terminology, the "transmitting system" and the "receiving system".

The second distortion affecting the voice spectrum is the attenuation of the subscriber lines. In a simple model of the local analogue line (given in a CNET Technical Note NT/LAA/ELR/289 by Cadoret, 1983), it is considered that this introduces an attenuation of the signal whose value in dB depends on its length and is proportional to the square root of the frequency. The attenuation is 3 dB at 800 Hz for an average line (approximately 2 km), 9.5 dB at 800 Hz for longer lines (up to 10 km). According to this model, the expression for the attenuation of a line, depicted in Figure 4, is:

$$A_{dB}(f) = A_{dB}(800Hz)\sqrt{\frac{f}{800}} \qquad (0.1)$$

To these distortions there is added the anti-aliasing filtering of the MIC coder (ref 30). The latter is typically a 200-3400 Hz bandpass filter with a response which is almost flat over the bandwidth and high attenuation outside the band, according to the template in Figure 5 for example (National Semiconductor, August 1994: Technical Documentation TP3054, TP3057).

Finally, the voice suffers spectral distortion as depicted in Figure 6 for the various combinations of three types of analogue line in transmission and reception (that is to say 6 distortions), assuming equipment complying with the nominal characteristic of the modified SRI. The voice thus appears to be stifled if one of the analogue lines is long and in all cases suffers from a lack of "presence" due to the attenuation of the low-frequency components.

1.2. Degradations in the timbre of the voice on the ISDN network and the GSM mobile network

In ISDN and the GSM network, the signal is digitised as from the terminal. The only analogue parts are the transmission and reception transducers associated with their respective amplification and conditioning chains. The UIT-T has defined frequency efficacy templates for transmission depicted in Figure 7, and for reception depicted in Figure 8, valid both for cabled digital telephones (UIT-T, Recommendation P.310, May 2000) and mobile digital or wireless terminals (UIT-T, Recommendation P.313, September 1999).

Moreover, for GSM networks, it is recognised that coding and decoding slightly modify the spectral envelope of the signal. This alteration is shown in Figure 9 for pink noise coded and then decoded in EFR (Enhanced Full Rate) mode.

The effect of these filterings on the timbre is mainly an attenuation of the low-frequency components, less marked however than in the case of STN.

The invention concerns the correction of these spectral distortions by means of a centralized ~~centralised~~ processing, that is to say a device installed in the digital part of the network, as indicated in Figure 10 for the STN.

The objective of a correction of the voice timbre is that the voice timbre in reception is as close as possible to that of the voice emitted by the speaker, which will be termed the original voice.

2. Prior art

Compensation for the spectral distortions introduced into the speech signal by the various elements of the telephone connection is at the present time allowed by devices with an equalization ~~equalisation~~ base. The latter can be fixed or be adapted according to the transmission conditions.

2.1. Fixed equalization ~~equalisation~~

Centralised equalization ~~equalisation~~ devices were proposed in the patents US 5333195 (Duane O. Bowker) and US 5471527 (Helena S. Ho). These equalizers ~~equaliser~~ are fixed filters which restore the level of the low frequencies attenuated by the transmitter. Bowker proposes for example a gain of 10 to 15 dB on the 100-300 Hz band. These methods have two drawbacks:

* The equalizer ~~equaliser~~ compensates only for the

filtering of the transmitter, so that on reception the low-frequency components remain greatly attenuated by the IRS reception filtering.

* This fixed equalization ~~equalisation~~ compensates for the average transmission conditions (transmission system and line). If the actual conditions are too different (for example if the analogue lines are long) the device does not sufficiently correct the timbre, or even impairs it more than the connection without equalization ~~equalisation~~.

2.2. Adaptive equalization ~~equalisation~~

The invention described in the patent US 5915235 (Andrew P De Jaco) aims to correct the non-ideal frequency response of a mobile telephone transducer. The equalizer ~~equaliser~~ is described as being placed between the analogue to digital converter and the CELP coder but can be equally well in the terminal or in the network. The principle of equalization ~~equalisation~~ is to bring the spectrum of the received signal close to an ideal spectrum. Two methods are proposed.

The first method (illustrated by Figure 4 in the aforementioned patent of De Jaco) consists of calculating long-term autocorrelation coefficients R_{LT}:

$$R_{LT}(n,i) = \alpha\, R_{LT}(n-1,i) + (1-\alpha)R(n,i), \qquad (0.2)$$

with $R_{LT}(n,i)$ the i^{th} long-term autocorrelation coefficient to the n^{th} frame, $R(n,i)$ the i^{th} autocorrelation coefficient specific to the n^{th} frame, and α a smoothing constant fixed for example at 0.995. From these coefficients there are derived the long-term LPC coefficients, which are the coefficients of a whitening filter. At the output of this filter, the signal is filtered by a fixed signal which imprints on it the ideal long-term spectral characteristics, i.e. those which it would have at the output of a transducer having the ideal frequency response. These two filters are supplemented by a multiplicative gain equal to the ratio between the long-term energies of the input of the whitener and the output of the second filter.

The second method, illustrated by Figure 5 of the aforementioned De Jaco patent, consists of dividing the signal into sub-bands and, for each sub-band, applying a multiplicative gain so as to reach a target energy, this gain being defined as the ratio between the target energy of the sub-band and the long-term energy (obtained by a smoothing of the instantaneous energy) of the signal in this sub-band.

These two methods have the drawback of correcting only the non-ideal response of the transmission system and not that of the reception system.

The object of the device of the patent US 5905969 (Chafik Mokbel) is to compensate for the filtering of

the transmission signal and of the subscriber line in order to improve the centralised recognition of the speech and/or the quality of the speech transmitted. As presented by Figure 3a in Mokbel, the spectrum of the signal is divided into 24 sub-bands and each sub-band energy is multiplied by an adaptive gain. The matching of the gain is achieved according to the stochastic gradient algorithm, by minimisation of the square error, the error being defined as the difference between the sub-band energy and a reference energy defined for each sub-band. The reference energy is modulated for each frame by the energy of the current frame, so as to respect the natural short-term variations in level of the speech signal. The convergence of the algorithm makes it possible to obtain as an output the 24 equalized ~~equalised~~ d sub-band signals.

If the application aimed at is the improvement in the voice quality, the equalized ~~equalised~~ speech signal is obtained by inverse Fourier transform of the equalized ~~equalised~~ sub-band energy.

The Mokbel patent does not mention any results in terms of improvement in the voice quality, and recognises that the method is sub-optimal, in that it uses a circular convolution. Moreover, it is doubtful that a speech signal can be reconstructed correctly by the inverse Fourier transform of band energies distributed according to the MEL scale. Finally, the

device described as not correct the filtering of the reception signal and of the analogue reception line.

The compensation for the line effect is achieved in the "Mokbel" method of cepstral subtraction, for the purpose of improving the robustness of the speech recognition. It is shown that the cepstrum of the transmission channel can be estimated by means of the mean cepstrum of the signal received, the latter first being whitened by a pre-accentuation filter. This method affords a clear improvement in the performance of the recognition systems but is considered to be an "off-line" method, 2 to 4 seconds being necessary for estimating the mean cepstrum.

2.3. Another state of the art combines a fixed pre-equalization ~~pre-equalisation~~ with an adapted equalization ~~equalisation~~ and has been the subject of the filing of a patent application FR 2822999 by the applicant. The device described aims to correct the timbre of the voice by combining two filters.

A fixed filter, called the pre-equalizer ~~pre-equaliser~~, compensates for the distortions of an average telephone line, defined as consisting of two average subscriber lines and transmission and reception systems complying with the nominal frequency responses defined in UIT-T, Recommendation P.48, App.I, 1988. Its frequency response on the Fc-3150 Hz band is the inverse of the global response of the analogue part of

this average connection, Fc being the limit equalization ~~equalisation~~ low frequency.

This pre-equalization ~~pre-equalisation~~ is supplemented by an adapted equalizer ~~equaliser~~, which adapts the correction more precisely to the actual transmission conditions. The frequency response of the adapted equalizer ~~equaliser~~ is given by:

$$|EQ(f)| = \frac{1}{|S\_RX(f).L\_RX(f)|}\sqrt{\frac{\gamma_{ref}(f)}{\gamma_x(f)}}, \qquad (0.3)$$

with L_RX the frequency response of the reception line, S_RX the frequency response of the reception system and $\gamma_x(f)$ the long-term spectrum of the output x of the pre-equalizer ~~pre-equaliser~~.

The long-term spectrum is defined by the temporal mean of the short-term spectra of the successive frames of the signal; $\gamma_{ref(f)}$, referred to as the reference spectrum, is the mean spectrum of the speech defined by the UIT (UIT-T/P.50/App. I, 1998), taken as an approximation of the original long-term spectrum of the speaker. Because of this approximation, the frequency response of the adapted equalizer ~~equaliser~~ is very irregular and only its general shape is pertinent. This is why it must be smoothed. The adapted equalizer ~~equaliser~~ being produced in the form of a time filter RIF, this smoothing in the frequency domain is obtained by a narrow windowing (symmetrical) of the pulsed

response.

This method makes it possible to restore a timbre close to that of the original signal on the equalization ~~equalisation~~ band (Fc-3150 Hz), but:

- for some speakers, the approximation of their original long-term spectrum by means of the reference spectrum is very rough, so that the equalizer ~~equaliser~~ introduces a perceptible distortion;

- the high smoothing of the frequency response of the equalizer ~~equaliser~~, made necessary by the approximation error, prevents fine spectral distortions from being corrected.

SUMMARY OF THE INVENTION

The aim of the invention is to remedy the drawbacks of the prior art. Its object is a method and system for improving the correction of the timbre by reducing the approximation error in the original long-term spectrum of the speakers.

To this end, it is proposed to classify the speakers according to their long-term spectrum and to approximate this not by a single reference spectrum but by one reference spectrum per class. The method proposed makes it possible to carry out an equalization ~~equalisation~~ processing able to determine the class of the speaker and to equalize ~~equalise~~ according to the

reference spectrum of the class. This reduction in the approximation error makes it possible to smooth the frequency response of the adapted equalizer ~~equaliser~~ less strongly, making it able to correct finer spectral distortions.

The object of the present invention is more particularly a method of correcting spectral deformations in the voice, introduced by a communication network, comprising an operation of equalization ~~equalisation~~ on a frequency band (F1-F2), adapted to the actual distortion of the transmission chain, this operation being performed by means of a digital filter having a frequency response which is a function of the ratio between a reference spectrum and a spectrum corresponding to the long-term spectrum of the voice signal of the speakers, principally characterised in that it comprises:

* prior to the operation of equalization ~~equalisation~~ of the voice signal of a speaker communicating:

- the constitution of classes of speakers with one voice reference per class,

* then, for a given speaker communicating:

- the classification of this speaker, that is to say his allocation to a class from predefined classification criteria in order to make a voice reference which is closest to his own correspond to him,

- the equalization ~~equalisation~~ of the digitised signal of the voice of the speaker carried out with, as

a reference spectrum, the voice reference of the class to which the said speaker has been allocated.

According to another characteristic, the constitution of classes of speakers comprises:

- the choice of a corpus of N speakers recorded under non-degraded conditions and the determination of their long-term frequency spectrum,

- the classification of the speakers in the corpus according to their partial cepstrum, that is to say the cepstrum calculated from the long-term spectrum restricted to the equalization ~~equalisation~~ band (F1-F2) and applying a predefined classification criterion to these cepstra in order to obtain K classes,

- the calculation of the reference spectrum associated with each class so as to obtain a voice reference corresponding to each of the classes.

According to another characteristic, the reference spectrum on the equalization ~~equalisation~~ frequency band (F1-F2), associated with each class, is calculated by Fourier transform of the ~~centre~~ center of the class defined by its partial cepstrum.

According to another characteristic, the classification of a speaker comprises:

- use of the mean pitch of the voice signal and of the partial cepstrum of this signal as classification parameters,

- the application of a discriminating function to these parameters in order to classify the said speaker.

According to the invention the method also

comprises a step of pre-equalization ~~pre-equalisation~~ of the digital signal by a fixed filter having a frequency response in the frequency band (F1-F2), corresponding to the inverse of a reference spectral deformation introduced by the telephone connection.

According to another characteristic, the equalization ~~equalisation~~ of the digitised signal of the voice of a speaker comprises:

- the detection of a voice activity on the line in order to trigger a concatenation of processings comprising the calculation of the long-term spectrum, the classification of the speaker, the calculation of the modulus of the frequency response of the equalizer ~~equaliser~~ filter restricted to the equalization ~~equalisation~~ band (F1-F2) and the calculation of the coefficients of the digital filter differentiated according to the class of the speaker, from this modulus,

- the control of the filter with the coefficients obtained,

- the filtering of the signal emerging from the pre-equalizer ~~pre-equaliser~~ by the said filter.

According to another characteristic, the calculation of the modulus (EQ) of the frequency response of the equalizer ~~equaliser~~ filter restricted to the equalization ~~equalisation~~ band (F1-F2) is achieved by the use of the following equation:

$$|EQ(f)| = \frac{1}{|S\_RX(f).L\_RX(f)|}\sqrt{\frac{\gamma_{ref}(f)}{\gamma_x(f)}}, \quad (0.3)$$

in which $\gamma_{ref}(f)$ is the reference spectrum of the class to which the said speaker belongs,

and in which L_RX is the frequency response of the reception line, S_RX is the frequency response of the reception signal and $\gamma_x(f)$ the long-term spectrum of the input signal x of the filter.

According to a variant, the calculation of the modulus of the frequency response of the equalizer ~~equaliser~~ filter restricted to the equalization ~~equalisation~~ band (F1-F2) is done using the following equation:

$$C^p_{eq} = C^p_{ref} - C^p_x - C^p_{S\_RX} - C^p_{L\_RX}, \qquad (0.13)$$

in which C^p_{eq}, C^p_x, $C^p_{S\_RX}$ and $C^p_{L\_RX}$ are the respective partial cepstra of the adapted equalizer ~~equaliser~~, of the input signal x of the equalizer ~~equaliser~~ filter, of the reception system and of the reception line, C^p_{ref} being the reference partial cepstrum, the ~~centre~~ center of the class of the speaker. The modulus (EQ) restricted to the band F1-F2 is then calculated by discrete Fourier transform of C^p_{eq}.

Another object of the invention is a system for correcting voice spectral deformations introduced by a communication network, comprising adapted equalization ~~equalisation~~ means in a frequency band (F1-F2) which comprise a digital filter whose frequency response is a

function of the ratio between a reference spectrum and a spectrum corresponding to the long-term spectrum of a voice signal, principally characterised in that these means also comprise:

- means of processing the signal for calculating the coefficients of the digital signal provided with:

- a signal processing unit for calculating the modulus of the frequency response of the equalizer ~~equaliser~~ filter restricted to the equalization ~~equalisation~~ band (F1-F2) according to the following equation:

$$|EQ(f)| = \frac{1}{|S\_RX(f).L\_RX(f)|}\sqrt{\frac{\gamma_{ref}(f)}{\gamma_x(f)}}, \qquad (0.3)$$

in which $\gamma_{ref}(f)$ is the reference spectrum, which may be different from one speaker to another and which corresponds to a reference for a predetermined class to which the said speaker belongs, and in which L_RX is the frequency response of the reception line, S_RX the frequency response of the reception signal and $\gamma_x(f)$ the long-term spectrum of the input signal x of the filter;

- a second processing unit for calculating the pulsed response from the frequency response modulus thus calculated, in order to determine the coefficients of the filter differentiated according to the class of the speaker.

According to another characteristic, the first processing unit comprises means of calculating the partial cepstrum of the equalizer ~~equaliser~~ filter according to the equation:

$$C_{eq}^{p} = C_{ref}^{p} - C_{x}^{p} - C_{S\_RX}^{p} - C_{L\_RX}^{p}\,, \qquad (0.13)$$

in which C_{eq}^{p}, C_{x}^{p}, $C_{S\_RX}^{p}$ and $C_{L\_RX}^{p}$ are the respective partial cepstra of the adapted equalizer ~~equaliser~~, of the input signal x of the equalizer ~~equaliser~~ filter, of the reception signal and of the reception line, C_{ref}^{p} being the reference partial cepstrum, the ~~centre~~ center of the class of the speaker, the modulus of (EQ) restricted to the band F1-F2 is then calculated by discrete Fourier transform of C_{eq}^{p}.

According to another characteristic, the first processing unit comprises a sub-assembly for calculating the coefficients of the partial cepstrum of a speaker communicating and a second sub-assembly for effecting the classification of this speaker, this second sub-assembly comprising a unit for calculating the pitch F_0, a unit for estimating the mean pitch from the calculated pitch F_0, and a classification unit applying a discriminating function to the vector x having as its components the mean pitch and the coefficients of the partial cepstrum for classifying the said speaker.

According to the invention, the system also

comprises a pre-equalization ~~pre-equalisation~~, the signal equalized ~~equalised~~ from reference spectra differentiated according to the class of the speaker being the output signal x of the pre-equalizer ~~pre-equaliser~~.

BRIEF DESCRIPTION OF THE DRAWINGS

Other particularities and advantages of the invention will emerge clearly from the following description, which is given by way of illustrative and non-limiting example and which is made with regard to the accompanying figures, which show:

- Figure 1, a diagrammatic telephone connection for a switched telephone network (STN),
- Figure 2, the transmission frequency response curve of the modified intermediate reference system IRS,
- Figure 3, the reception frequency response curve of the modified intermediate reference system IRS,
- Figure 4, the frequency response of the subscriber lines according to their length,
- Figure 5, the template of the anti-aliasing filter of the MIC coder,
- Figure 6, the spectral distortions suffered by the speech on the switched telephone network with average IRS and various combinations of analogue lines,
- Figure 7, the transmission template for the digital terminals,
- Figure 8, the reception template for the digital

terminals,

- Figure 9, the spectral distortion introduced by GSM coding/decoding in EFR (Enhanced Full Rate) mode,

- Figure 10, the diagram of a communication network with a system for correcting the speech distortions,

- Figure 11, the steps of calculating the partial cepstrum,

- Figure 12, the classification of the partial cepstra according to the variance criterion,

- Figures 13a and 13b, the long-term spectra corresponding to the ~~centres~~ centers of the classes of speakers respectively for men and women,

- Figure 14, the frequency characteristics of the filterings applied to the corpus in order to define the learning corpus,

- Figure 15, the frequency response of the pre-equalizer ~~pre-equaliser~~ for various frequencies Fc,

- Figure 16, the scheme for implementing the system of correction by differentiated equalization ~~equalisation~~ per class of speaker,

- Figure 17, a variant execution of the system according to Figure 16.

DETAILED DESCRIPTION OF THE DRAWINGS

Throughout the following the same references entered on the drawings correspond to the same elements.

The description which follows will first of all present the prior step of classification of a corpus of speakers according to their long-term spectrum. This step defines K classes and one reference per class.

A concatenation of processings makes it possible to process the speech signal (as soon as a voice activity is detected by the system) for each speaker in order on the one hand to classify the speakers, that is to say to allocate them to a class according to predetermined criteria, and on the other hand to correct the voice using the reference of the class of the speaker.

Prior step of classification of the speakers.

* Choice of the class definition corpus.

The reference spectrum being an approximation of the original long-term spectrum of the speakers, the definition of the classes of speakers and their respective reference spectra requires having available a corpus of speakers recorded under non-degraded conditions. In particular, the long-term spectrum of a speaker measured on this recording must be able to be considered to be its original spectrum, i.e. that of its voice at the transmission end of a telephone connection.

Definition of the individual: the partial cepstrum

The processing proposed makes it possible to have available, in each class, a reference spectrum as close as possible to the long-term spectrum of each member of the class. However, only the part of the spectrum included in the ~~equalisation~~ equalization band F1-F2 is taken into account in the adapted ~~equalisation~~ equalization processing. The classes are therefore formed according to the long-term spectrum restricted to this band.

Moreover, the comparison between two spectra is made at a low spectral resolution level, so as to reflect only the spectral envelope. This is why the space of the first cepstral coefficients of order greater than 0 (the coefficient of order 0 representing the energy) is preferably used, the choice of the number of coefficients depending on the required spectral resolution.

The "long-term partial cepstrum", which is denoted Cp, is then determined in the processing as the cepstral representation of the long-term spectrum restricted to a frequency band. If the frequency indices corresponding respectively to the frequencies F1 and F2 are denoted k1 and k2 and the long-term spectrum of the speech is denoted γ, the partial cepstrum is defined by the equation:

$$C^{p} = TFD^{-1}\left(10\log\left(\gamma\left(k_1 \ldots k_2\right) \circ \gamma\left(k_2 - 1 \ldots k_1 + 1\right)\right)\right) \qquad (0.4)$$

where o designates the concatenation operation.

The inverse discrete Fourier transform is calculated for example by IFFT after interpolation of the samples of the truncated spectrum so as to achieve a number of power samples of 2. For example, by choosing the ~~equalisation~~ equalization band 187-3187 Hz, corresponding to the frequency indices 5 to 101 for a representation of the spectrum (made symmetrical) on 256 points (from 0 to 255) the interpolation is made simply by interposing a frequency line (interpolated linearly) every three lines in the spectrum restricted to 187-3187 Hz.

The steps of the calculation of the partial cepstrum are shown in Figure 11.

For the cepstral coefficients to reflect the spectral envelope but not the influence of the harmonic structure of the spectrum of the speech on the long-term spectra, the high-order coefficients are not kept. The speakers to be classified are therefore represented by the coefficients of orders 1 to L of their long-term partial cepstrum, L typically being equal to 20.

* The classification.

The classes are formed for example in a non-

supervised manner, according to an ascending hierarchical classification.

This consists of creating, from N separate individuals, a hierarchy of partitionings according to the following process: at each step, the two closest elements are aggregated, an element being either a non-aggregated individual or an aggregate of individuals formed during a previous step. The proximity between two elements is determined by a measurement of dissimilarity which is called distance. The process continues until the whole population is aggregated. The hierarchy of partitionings thus created can be represented in the form of a tree like the one in Figure 12, containing N-1 imbricated partitionings. Each cut of the tree supplies a partitioning, which is all the finer, the lower the cut.

In this type of classification, as a measurement of distance between two elements, the intra-class inertia variation resulting from their aggregation is chosen. A partitioning is in fact all the better, the more homogeneous are the classes created, that is to say the lower the intra-class inertia. In the case of a cloud of points xi with respective masses mi, distributed in q classes with respective ~~centres~~ <u>centers</u> of gravity gq, the intra-class inertia is defined by:

$$I_{intra} = \sum_q \sum_{i \in q} m_i \left\| x_i - g_q \right\|^2 . \qquad (0.5)$$

The intra-class inertia, zero at the initial step of the calculation algorithm, inevitably increases with each aggregation.

Use is preferably made of the known principle of aggregation according to variance. According to this principle, at each step of the algorithm used, the two elements are sought whose aggregation produces the lowest increase in intra-class inertia.

The partitioning thus obtained is improved by a procedure of aggregation around the movable ~~centres~~ centers, which reduces the intra-class variance.

The reference spectrum, on the band F1-F2, associated with each class is calculated by Fourier transform of the ~~centre~~ center of the class.

* Example of classification.

The processing described above is applied to a corpus of 63 speakers. The classification tree of the corpus is shown in Figure 12. In this representation, the height of a horizontal segment aggregating two elements is chosen so as to be proportional to their distance, which makes it possible to display the proximity of the elements grouped together in the same

class. This representation facilitates the choice of the level of cutoff of the tree and therefore of the classes adopted. The cutoff must be made above the low-level aggregations, which group together close individuals, and below the high-level aggregations, which associate clearly distinct groups of individuals.

In this way, four classes are clearly obtained (K = 4). These classes are very homogeneous from the point of view of the sex of the speakers, and a division of the tree into two classes shows approximately one class of men and one class of women.

The consolidation of this partitioning by means of an aggregation procedure around the movable ~~centres~~ centers results in four classes of cardinals 11, 18, 18 and 16, more homogeneous than before from the point of view of the sex: only one man and two women are allocated to classes not corresponding to their sex.

The spectra restricted to the 187-3187 Hz band corresponding to the ~~centres~~ centers of these classes are shown in Figures 13a and 13b for the men and women classes as well as for their respective sub-classes. These spectra, the results of the classification, are used as a multiple reference by the adapted equalizer ~~equaliser~~.

* Use of classification criteria for the speakers

The classes of speakers being defined, the processing provides for the use of parameters and criteria for allocating a speaker to one or other of the classes.

This allocation is not carried out simply according to the proximity of the partial cepstrum with one of the class ~~centres~~ centers, since this cepstrum is diverted by the part of the telephone connection upstream of the equalizer ~~equaliser~~.

It is advantageously proposed to use classification criteria which are robust to this diversion. This robustness is ensured both by the choice of the classification parameters and by that of the classification criteria learning corpus.

* Preferably the classification parameters average pitch and partial cepstrum are used

The classes previously defined are homogeneous from the point of view of the sex. The average pitch being both fairly discriminating for a man/woman classification and insensitive to the spectral distortions caused by a telephone connection, and is therefore used as a classification parameter conjointly with the partial cepstrum.

* Choice of the classification criteria learning corpus

A discrimination technique is applied to these parameters, for example the usual technique of discriminating linear analysis.

Other known techniques can be used such as a non-linear technique using a neural network.

If N individuals are available, described by dimension vectors p and distributed a priori in K classes, the discriminating linear analysis consists of:

- firstly, seeking the K-1 independent linear functions which best separate the K classes. It is a case of determining which are the linear combinations of the p components of the vectors which minimise the intra-class variance and maximise the inter-class variance;

- secondly, determining the class of a new individual by applying the discriminating linear functions to the vector representing him.

In the present case, the vectors representing the individuals have as their components the pitch and the coefficients 1 to L (typically L = 20) of the partial cepstrum. The robustness of the discriminating functions to the deviation of the cepstral coefficients is ensured both by the presence of the pitch in the

parameters and by the choice of the learning corpus. The latter is composed of individuals whose original voice has undergone a great diversity of filtering representing distortions caused by the telephone connections.

More precisely, from a corpus of original voices (non-degraded) of N speakers, there is defined a corpus of N vectors of components $\left|\overline{F_0};C^p(1);...;C^p(L)\right|$, with $\overline{F_0}$ the mean pitch and C^p the partial cepstrum. The construction of the learning corpus of the said functions consists of defining a set of M cepstral biases which are each added to each partial cepstrum representing a speaker in the original corpus, which makes it possible to obtain a new corpus of NM individuals.

These biases in the domain of the partial cepstrum correspond to a wide range of spectral distortions of the band F1-F2, close to those which may result from the telephone connection.

By way of example, the set of frequency responses depicted in Figure 14 is proposed for the 187-3187 Hz band: each frequency response corresponds to a path from left to right in the lattice. The amplitude of their variations on this band does not exceed 20 dB, like extreme characteristics of the transmission and line systems.

From these 81 frequency characteristics there are

calculated the 81 corresponding biases in the domain of the partial cepstrum, according to the processing described for the use of equation (0.4). By the addition of these biases to the corpus of 63 speakers previously used, a learning corpus is obtained including 5103 individuals representing various conditions (speaker, filtering of the connection).

In the case of classification by discriminating linear analysis:

* Application of the classification critéria

Let $(a^k)1 \leq k \leq K-1$ be the family of discriminating linear functions defined from the learning corpus. A speaker represented by the vector $x = \left|\overline{F}_0; C^p(1); \ldots; C^p(L)\right|$ is allocated to the class q if the conditional probability of q knowing a(x), denoted $P(q|a(x))$, is maximum, a(x) designating the vector of components $(a^k(x))1 \leq k \leq K-1$. According to Bayes' theorem,

$$P(q \mid a(x)) = \frac{P(a(x) \mid q) P(q)}{P(a(x))}. \qquad (0.6)$$

Consequently $P(q|a(x))$ is proportional to $P(a(x)|q)P(q)$. In the subspace generated by the K-1 discriminating functions, on the assumption of a multi-Gaussian distribution of the individuals in each class, the density of probability of a(x) within the class q has:

$$f_q(x)=\frac{1}{(2\pi)^{\frac{K-1}{2}}\sqrt{|S_q|}}exp\left(-\frac{1}{2}\left(a(x)-a\left(\overline{x}^q\right)\right)' S_q^{-1}\left(a(x)-a\left(\overline{x}^q\right)\right)\right),$$

(0.7)

where $\overline{x}^q$ is the ~~centre~~ center of the class q, |Sq| designates the determinant of the matrix Sq, and Sq is the matrix of the covariances of a within the class q, of generic element $\sigma^q jk$, which can be estimated by:

$$\sigma^q_{jk}=\frac{1}{N_q}\sum_{j=1}^{N_q}\left(a^j\left(x^i\right)-a^j\left(\overline{x}^q\right)\right)\left(a^k\left(x^i\right)-a^k\left(\overline{x}^q\right)\right). \quad (0.8)$$

The individual x will be allocated to the class q which maximises fq(x)P(q), which amounts to minimising on q the function sq(x) also referred to as the discriminating score:

$$s_q(x)=\left(a(x)-a\left(\overline{x}^q\right)\right)' S_q^{-1}\left(a(x)-a\left(\overline{x}^q\right)\right)+log(|S_q|)-2log(P(q)),$$

(0.9)

The correction method proposed is implemented by the correction system (equalizer ~~equaliser~~) located in the digital network 40 as illustrated in Figure 10.

Figure 16 illustrates the correction system able to implement the method. Figure 17 illustrates this system according to a variant embodiment as will be

detailed hereinafter. These variants relate to the method of calculating the modulus of the frequency response of the adapted equalizer ~~equaliser~~ restricted to the band F1-F2.

The pre-equalizer ~~pre-equaliser~~ 200 is a fixed filter whose frequency response, on the band F1-F2, is the inverse of the global response of the analogue part of an average connection as defined previously (UIT-T/P.830, 1996).

The stiffness of the frequency response of this filter implies a long-pulsed response; this is why, so as to limit the delay introduced by the processing, the pre-equalizer ~~pre-equaliser~~ is typically produced in the form of an RII filter, 20^{th} order for example.

Figure 15 shows the typical frequency responses of the pre-equalizer ~~pre-equaliser~~ for three values of F1. The scattering of the group delays is less than 2 ms, so that the resulting phase distortion is not perceptible.

The processing chain 400 which follows allows classification of the speaker and differentiated matched equalization ~~equalisation~~. This chain comprises two processing units 400A and 400B. The unit 400A makes it possible to calculate the modulus of the frequency response of the equalizer ~~equaliser~~ filter restricted to the equalization ~~equalisation~~ band: EQ dB (F1-F2).

The second unit 400B makes it possible to calculate the pulsed response of the equalizer ~~equaliser~~ filter in order to obtain the coefficients eq(n) of the differentiated filter according to the class of the speaker.

A voice activity frame detector 401 triggers the various processings.

The processing unit 410 allows classification of the speaker.

The processing unit 420 calculates the long-term spectrum followed by the calculation of the partial cepstrum of this speaker.

The output of these two units is applied to the operator 428a or 428b. The output of this operator supplies the modulus of the frequency response of the equalizer ~~equaliser~~ matched for dB restricted to the equalization ~~equalisation~~ band F1-F2 via the unit 429 for 428a, via the unit 440 for 428b.

The processing units 430 to 435 calculate the coefficients eq(n) of the filter.

The output x(n) of the pre-equalizer ~~pre-equaliser~~ is analysed by successive frames with a typical duration of 32 ms, with an interframe overlap of

typically 50%. For this purpose an analysis window represented by the blocks 402 and 403 is opened.

The matched equalization ~~equalisation~~ operation is implemented by an RIF filter 300 whose coefficients are calculated at each voice activity frame by the processing chain illustrated in Figures 16 and 17.

The calculation of these coefficients corresponds to the calculation of the pulsed response of the filter from the modulus of the frequency response.

The long-term spectrum of x(n), γ_x, is first of all calculated (as from the initial moment of functioning) on a time window increasing from 0 to a voice activity duration T (typically 4 seconds), and then adjusted recursively to each voice activity frame, which is represented by the following generic formula:

$$\gamma_x(f,n) = \alpha(n)|X(f,n)|^2 + (1-\alpha(n))\gamma_x(f,n-1),$$
(0.10)

where γ_x (f,n) is the long-term spectrum of x at the n^{th} voice activity frame, X(f,n) the Fourier transform of the n^{th} voice activity frame, and $\alpha(n)$ is defined by equation (0.11). Denoting N the number of frames in the period T,

$$\alpha(n) = \frac{1}{min(n,N)} . \qquad (0.11)$$

This calculation is carried out by the units 421, 422, 423.

Next there is calculated, from this long-term spectrum, the partial cepstrum Cp, according to the equation (0.4), used by the processing units 424, 425, 426.

The mean pitch $\overline{F_0}$ is estimated by the processing unit 412 at each voiced frame according to the formula:

$$\overline{F_0}(m) = \alpha(m)F_0(m) + (1-\alpha(m))\overline{F_0}(m-1) , \qquad (0.12)$$

where F0(m) is the pitch of the m^{th} voiced frame and is calculated by the unit 411 according to an appropriate method of the prior art (for example the autocorrelation method, with determination of the voicing by comparison of the standardized ~~standardised~~ autocorrelation with a threshold (UIT-T/G.729, 1996).

Thus, at each voice activity frame, there is a new vector x of components, the mean pitch and the coefficients 1 to L of the partial cepstrum, to which there is applied the discriminating function a defined from the learning corpus. This processing is implemented by the unit 413. The speaker is then allocated to the minimum discriminating score class q.

The modulus in dB of the frequency response of the matched equalizer ~~equaliser~~ restricted to the band F1-F2, denoted $|EQ|_{dB(F1-F2)}$, is calculated according to one of the following two methods:

The first method (Figure 16) consists of calculating $|EQ|_{F1-F2}$ according to equation (0.3), where $\gamma_{ref}(f)$ is the reference spectrum of the class of the speaker (Fourier transform of the class center ~~centre~~). This calculation method is implemented in this variant depicted in Figure 16 with the operators 414a, 428a, 427 and 429.

The second method (Figure 17) consists of transcribing equation (0.3) into the domain of the partial cepstrum, and then the partial cepstrum of the output x of the pre-equalization ~~pre-equalisation~~, necessary for the classification of the speaker, is available. Thus equation (0.3) becomes:

$$C_{eq}^{p} = C_{ref}^{p} - C_{x}^{p} - C_{S\_RX}^{p} - C_{L\_RX}^{p}, \qquad (0.13)$$

where C_{eq}^{p}, C_{x}^{p}, $C_{S\_RX}^{p}$ and $C_{L\_RX}^{p}$ are the respective partial cepstra of the matched equalizer ~~equaliser~~, of the output x of the pre-equalizer ~~pre-equaliser~~, of the reception system and of the reception line, C_{ref}^{p} being the reference partial cepstrum, the center ~~centre~~ of the class of the speaker. The partial cepstra are

calculated as indicated before, selecting the frequency band F1-F2. This calculation is made solely for the coefficients 1 to 20, the following coefficients being unnecessary since they represent a spectral fineness which will be eliminated subsequently.

The 20 coefficients of the partial cepstrum of the matched equalizer ~~equaliser~~ are obtained by the operators 414b and 428b according to equation (0.13).

The processing unit 441 supplements these 20 coefficients with zeros, makes them symmetrical and calculates, from the vector thus formed, the modulus in dB of the frequency response of the matched equalizer ~~equaliser~~ restricted to the band F1-F2 using the following equation:

$$EQ_{dB(F_1-F_2)} = TFD^{-1}\left(C_{eq}^{p}\right). \qquad (0.14)$$

This response is decimated by a factor of ¾ by the operator 442.

For the two variants which have just been described, the values of $|EQ|$ outside the band F1-F2 are calculated by linear extrapolation of the value in dB of $|EQ|_{F1-F2}$, denoted EQ_{dB} hereinafter, by the unit 430 in the following manner:

For each index of frequency k, the linear approximation of EQ_{dB} is expressed by:

$$EQ_{dB}(k) = \alpha_1 + \alpha_2 k \qquad (0.15)$$

The coefficients a1 and a2 are chosen so as to minimise the square error of the approximation on the range F1-F2, defined by

$$e = \sum_{k-k_1}^{k_1} \left(EQ_{dB}(k) - EQ_{dB}(k)\right)^2 \qquad (0.16)$$

The coefficients a1 and a2 are therefore defined by:

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} k_2 - k_1 + 1 & \sum_{k-k_1}^{k_1} k \\ \sum_{k=k_1}^{k_2} k & \sum_{k=k_1}^{k_2} k^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{k=k_1}^{k_2} EQ_{dB}(k) \\ \sum_{k=k_1}^{k_2} kEQ_{dB}(k) \end{pmatrix} \qquad (0.17)$$

The values of |EQ|, in dB, outside the band F1-F2, are then calculated from the formula (0.15).

The frequency characteristic thus obtained must be smoothed. The filtering being performed in the time domain, the means allowing this smoothing is to multiply by a narrow window the corresponding pulsed response.

The pulsed response is obtained by an IFFT operation applied to |EQ| carried out by the units 431

and 432 followed by a symmetrization ~~symmetrisation~~ performed by the processing unit 433, so as to obtain a linear-phase causal filter. The resulting pulsed response is multiplied, operator 435, by a time window 434. The window used is typically a Hamming window of length 31 ~~centred~~ centered on the peak of the pulsed response and is applied to the pulsed response by means of the operator 435.

ABSTRACT OF THE DISCLOSURE

A technique for correcting the voice spectral deformations introduced by a communication network. Prior to the operation of equalization ~~equalisation~~ of the voice signal of a speaker, the constitution of classes of speakers is communicated, with one voice reference per class. Then, for a given speaker, the classification of this speaker is communicated, that is to say his allocation to a class from predefined classification criteria in order to make a voice reference which is closest to his own correspond to him. Then, for that given speaker, communicating the equalization ~~equalisation~~ of the digitized ~~digitised~~ signal of the voice of the speaker carried out with, as a reference spectrum, the voice reference of the class to which the speaker has been allocated. This technique applies to the correction of the timbre of the voice in switched telephone networks, in ISDN networks and in mobile networks.

Attorney Docket # 5394-3 Patent

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of

Gael MAHE et al.

Serial No.: 10/723,851

Filed: November 25, 2003

For: Method and System of Correcting Spectral Deformations in the Voice, Introduced by a Communication Network

Examiner: Wozniak, James S.
Group Art: 2626

I hereby certify that this correspondence is being deposited with the United States Postal Service with sufficient postage as first class mail in an envelope addressed to: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450, on

September 6, 2007
(Date of Deposit)

Edward M. Weisz
Name of applicant, assignee or Registered Representative

Signature

September 6, 2007
Date of Signature

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Certification

SIR:

The undersigned attorney certifies that the attached substitute specification does not contain new matter. All amendments made are clearly shown in the attached redlined specification.

Respectfully submitted,
COHEN, PONTANI, LIEBERMAN & PAVANE LLP

By

Edward M. Weisz
Reg. No. 37,257
551 Fifth Avenue, Suite 1210
New York, New York 10176
(212) 687-2770

Dated: September 6, 2007

38767_1.DOC